ATTY. DOCKET NO. SAMS01-00266 PATENT

APPARATUS AND METHOD FOR MAINTAINING PACKET SEQUENCING IN A PARALLEL ROUTER

Inventor:

Jack C. Wybenga
2129 Stone Creek
Plano
Collin County
Texas 75075
United States citizen

Patricia K. Sturm
4424 Rushing Road
Dallas
Collin County
Texas 75287
United States citizen

Pradeep D. Samudra
6509 Crawley Drive
Plano
Collin County
Texas 75093
United States citizen

Assignee:

SAMSUNG ELECTRONICS CO., LTD.
416, Maetan-dong, Paldal-gu
Suwon-city, Kyungki-do
Republic of Korea

William A. Munck
John T. Mockler
Davis Munck, P.C.
Three Galleria Tower
13155 Noel Road, Suite 900
Dallas, Texas 75240
(972) 628-3600

APPARATUS AND METHOD FOR MAINTAINING PACKET SEQUENCING IN A PARALLEL ROUTER

TECHNICAL FIELD OF THE INVENTION

[001] The present invention relates to massively parallel routers and, more specifically, to a mechanism for maintaining packet sequencing in a parallel router.

BACKGROUND OF THE INVENTION

[002] The growth of the Internet during the 1990s was the driving force behind the three generations of Internet Protocol (IP) routers. The development is underway of fourth generation routers, which include more optics and a higher degree of parallelism. While Internet bandwidth is no longer growing at triple digit rates, it is still growing at a respectable 60-70% per year. This is a seemingly sustainable growth, which exceeds Moore's law by a significant margin.

[003] To match this demand, vendors are delivering faster IP core routers that implement optical interfaces and electronic switching matrices to accommodate the inexorable growth in Internet traffic. The next evolution of the core network has an IP layer at the edge of a circuit-switched optical layer based on wave division multiplexing (WDM) circuits with optical cross connects. However,

there is reason to be concerned about the ability of primarily electronic IP routers to keep pace with the bandwidth growth provided by the switched optical layer.

[004] A packet switch that is fully optical requires a technological evolution that is currently just a promise. Many of the enabling technologies are still in the stage of research and experimentation. So, while optical switching may be deployed in the future, it is not expected to come soon enough to handle nearer term bandwidth needs. Thus, there will be a gap between Internet bandwidth needs and the bandwidth capabilities of primarily electronic IP routers before 100% optical packet switches become practical. In the near term, switch routers have the option of being simpler, using more optics, and taking advantage of increased parallelism.

[005] However, the issue of maintaining packet sequencing in fourth generation routers is becoming more problematic. Most conventional high performance packet switches use input queuing and a non-blocking (e.g., crossbar) switch fabric. Thin input queues are arranged as virtual output queues (VOQs) to overcome head of line blocking and to enable high throughput rates. To simplify the task of memory management, a fixed-sized time slot is used. This requires the segmenting of incoming variable length packets into

fixed-sized cells. A centralized scheduler examines each slot of the VOQ to determine the configuration of the switch fabric for the next time slot. These switch fabrics generally have a scheduler-based forwarding mechanism. A hot standby mode switch fabric provides redundancy.

[006] There has been significant work in the area of parallel operating switch fabrics. However, there has been limited work on maintaining packet sequencing in systems using such parallel operating switch fabrics. In 2001, Iyer and McKeown suggested using a line buffer to reorder mis-sequenced data packets. In June 2002, Keslassy and McKeown proposed a "full frames first" (FFF) mechanism that eliminates the sequencing buffer by avoiding data packet mis-sequencing. The FFF mechanism uses a three-dimensional variant of the virtual output queue and a set of deterministic sequences that connect inputs to outputs to achieve this feat. However, all of the proposed solutions are complex and hardware intensive, thus increasing the cost of the routers and decreasing their reliability.

[007] Therefore, there is a need in the art for an improved Internet protocol (IP) router. In particular, there is a need for a massively parallel, distributed architecture router that is

capable of minimizing the occurrence of out-of-sequence transmission of data packets from the router.

SUMMARY OF THE INVENTION

[008] To address the above-discussed deficiencies of the prior art, it is a primary object of the present invention to provide a router for interconnecting N interfacing peripheral devices. According to an advantageous embodiment of the present invention, the router comprises: i) a first switch fabric; ii) a second switch fabric; and iii) a plurality of routing nodes coupled to the first and second switch fabrics, each of the routing nodes comprising an input-output processing (IOP) module capable of forwarding received data packets to other ones of the IOP modules via the first and second switch fabrics, wherein a first one of the IOP modules forwards received data packets directed to a second one of the IOP modules by alternating between the first and second switch fabrics for each sequential data packet directed to the second IOP module.

[009] According to one embodiment of the present invention, the first IOP module forwards received data packets directed to a third one of the IOP modules by alternating between the first and second switch fabrics for each sequential data packet directed to the third IOP module.

[010] According to another embodiment of the present invention, the alternate selection of the first and second switch fabrics for forwarding of data packets between the first and second IOP modules

is independent of the alternate selection of the first and second switch fabrics for forwarding of data packets between the first and third IOP modules.

[011] According to still another embodiment of the present invention, the second IOP module is capable of determining that a next expected data packet from the first IOP module was not received in an alternating manner from the first and second switch fabrics.

[012] According to yet another embodiment of the present invention, the second IOP module, in response to the determination that the next expected data packet from the first IOP module was not received in an alternating manner from the first and second switch fabrics, determines that one of the first and second switch fabrics is faulty and ceases forwarding data packets to the first IOP module via the faulty one of the first and second switch fabrics.

[013] According to a further embodiment of the present invention, the first IOP module is capable of determining that a next expected data packet from the second IOP module was not received from the faulty one of the first and second switch fabrics and, in response to the determination, the first IOP module ceases

forwarding data packets to the second IOP module via the faulty one of the first and second switch fabrics.

[014] This has outlined rather broadly several features of this disclosure so that those skilled in the art may better understand the DETAILED DESCRIPTION that follows. Additional features may be described later in this document. Those skilled in the art should appreciate that they may readily use the concepts and the specific embodiments disclosed as a basis for modifying or designing other structures for carrying out the same purposes of this disclosure. Those skilled in the art should also realize that such equivalent constructions do not depart from the spirit and scope of the invention in its broadest form.

[015] Before undertaking the DETAILED DESCRIPTION below, it may be advantageous to set forth definitions of certain words and phrases used throughout this patent document. The terms "include" and "comprise," as well as derivatives thereof, mean inclusion without limitation. The term "or" is inclusive, meaning and/or. The phrases "associated with" and "associated therewith," as well as derivatives thereof, may mean to include, be included within, interconnect with, contain, be contained within, connect to or with, couple to or with, be communicable with, cooperate with, interleave, juxtapose, be proximate to, be bound to or with, have,

have a property of, or the like. The term "controller" means any device, system, or part thereof that controls at least one operation. A controller may be implemented in hardware, firmware, or software, or a combination of at least two of the same. It should be noted that the functionality associated with any particular controller may be centralized or distributed, whether locally or remotely. Definitions for certain words and phrases are provided throughout this patent document, and those of ordinary skill in the art should understand that in many, if not most instances, such definitions apply to prior as well as future uses of such defined words and phrases.

BRIEF DESCRIPTION OF THE DRAWINGS

[016] For a more complete understanding of the present invention, and the advantages thereof, reference is now made to the following descriptions taken in conjunction with the accompanying drawings, wherein like numbers designate like objects, and in which:

[017] FIGURE 1 illustrates an exemplary distributed architecture router, which implements packet sequencing according to the principles of the present invention;

[018] FIGURE 2 illustrates the routing of data packets between IOP modules in the distributed architecture router according to one embodiment of the present invention; and

[019] FIGURE 3 illustrates the selection of switch fabrics for the transmission of data packets between pairs of IOP modules in the exemplary distributed architecture router according to one embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[020] FIGURES 1 through 3, discussed below, and the various embodiments used to describe the principles of the present invention in this patent document are by way of illustration only and should not be construed in any way to limit the scope of the invention. Those skilled in the art will understand that the principles of the present invention may be implemented in any suitably arranged distributed router.

[021] FIGURE 1 illustrates exemplary distributed architecture router 100, which implements packet sequencing according to the principles of the present invention. Distributed architecture router 100 provides scalability and high-performance using up to N independent routing nodes (RN), including exemplary routing nodes 110, 120, 130 and 140, connected by switch 150, which comprises a pair of high-speed switch fabrics 155a and 155b. Each routing node comprises an input-output processor (IOP) module, and one or more physical medium device (PMD) module. Exemplary RN 110 comprises PMD module 112 (labeled PMD-a), PMD module 114 (labeled PMD-b), and IOP module 116. RN 120 comprises PMD module 122 (labeled PMD-a), PMD module 124 (labeled PMD-b), and IOP module 126. RN 130 comprises PMD module 132 (labeled PMD-a), PMD module 134 (labeled PMD-b), and IOP module 136. Finally, exemplary RN 140 comprises

PMD module 142 (labeled PMD-a), PMD module 144 (labeled PMD-b), and IOP module 146.

[022] Each one of IOP modules 116, 126, 136 and 146 buffers incoming Internet protocol (IP) frames and MPLS frames from subnets or adjacent routers, such as router 190 and network 195. Additionally, each of IOP modules 116, 126, 136 and 146 classifies requested services, looks up destination addresses from frame headers, and forwards frames to the outbound IOP module. Moreover, each IOP module also maintains an internal routing table determined from routing protocol messages and provisioned static routes and computes the optimal data paths from the routing table. Each IOP module processes an incoming frame from one of its PMD modules. According to one embodiment of the present invention, each PMD module frames an incoming frame (or cell) from an IP network (or ATM switch) for processing in an IOP module and performs bus conversion functions.

[023] Each one of routing nodes 110, 120, 130, and 140, configured with an IOP module and PMD module(s) and linked by switch fabrics 155a and 155b, is essentially equivalent to a router by itself. Thus, distributed architecture router 100 can be considered a set of RN building blocks with high-speed links (i.e., switch fabrics 155a and 155b) connected to each block. Switch

fabrics 155a and 155b support frame switching between IOP modules. Switch processor (SWP) 160a and switch processor (SWP) 160b, located in switch fabrics 155a and 155b, respectively, support system management.

[024] FIGURE 2 illustrates the routing of data packets between exemplary IOP modules in distributed architecture router 100 according to one embodiment of the present invention. As stated above, there are N routing nodes (RNs) in router 100 and each node contains an IOP module. FIGURE 2 illustrates the interconnections between switch fabrics 155a and 155b and the N IOP modules, including exemplary IOP modules 201-204. IOP modules 201-204 are arbitrarily labeled IOP 1, IOP 2, IOP 3, and IOP N. Switch fabrics 155a and 155b are labeled Switch Fabric A and Switch Fabric B, respectively.

[025] It is noted that the embodiment of router 100 shown in FIGURES 1 and 2 contains only two switch fabrics. However, this is by way of illustration only and should not be construed to limit the scope of the present invention. In alternate embodiments of the present invention, router 100 may contain three, four or more switch fabrics. The switch fabrics of router 100 may be referred to hereafter as Switch Fabric A, Switch Fabric B, Switch Fabric C, Switch Fabric D, and so forth. However, in an advantageous

embodiment of the present invention, two switch fabrics are sufficient to provide a redundant configuration.

[026] Router 100 differs from conventional router architectures for two fundamental reasons. First, router 100 uses Ethernet technology that relinquishes any control of switch 150. Secondly, there is a requirement for redundancy. The parallel load shared switch fabric of router 100 pathologically mis-sequences data packets because router 100 permits variable length packets. The present invention resolves this problem by creating a relationship between each source-destination pair of IOP modules.

[027] According to the principles of the present invention, each IOP module maintains an index of the range equivalent to the number of redundant switch fabrics permitted, for each adjacent IOP. For example, router 100 may use a binary table of length 0-255 in an implementation that permits 256 IOP modules and has two switch fabrics. A given source IOP module sends data packets to the destination IOP module via the switch fabrics using a round-robin algorithm. Thus, if four switch fabrics are used (e.g., Switch Fabric A, Switch Fabric B, Switch Fabric C, and Switch Fabric D), then data packets are sent A B C D A B C D . . . from the source IOP module to the destination module. Thus, the IOP

module selects the switch fabric for the next data packet to be transmitted based on the destination IOP module.

[028] In the case of a two switch fabric router, the round-robin algorithm causes data packets sent from IOP j to IOP k to alternate between a primary switch fabric (i.e., Switch Fabric A) and a secondary switch fabric (i.e., Switch Fabric B), so that the switch fabrics are selected as ABABABA . . . Data packets sent to the same destination IOP module may be interspersed with data packets being sent to different destination IOP modules. However, the round robin sequence for each pair of source and destination IOP modules is maintained separately.

[029] For example, FIGURE 3 illustrates the selection of switch fabrics for the transmission of data packets between pairs of IOP modules in exemplary distributed architecture router 100 according to one embodiment of the present invention. In FIGURE 3, data packets are sent from a single source IOP module (i.e., IOP 3) to three different destination IOP modules (i.e., IOP 2, IOP 4, and IOP 5). The transmission of each data packet is listed as one of transactions 301-314. As FIGURE 3 shows, router 100 maintains a separate round robin algorithm for each destination IOP module.

[030] For example, transactions 301, 306, 308 and 309 are transmissions between IOP 3 and IOP 2. Switch Fabric A and Switch

Fabric B are selected in the order [ABAB . . .] for transactions 301, 306, 308 and 309, despite the interleaving of transactions 302-305 and 307. Similarly, transactions 302-305 are transmissions of data packets between IOP 3 and IOP 5. Switch Fabric A and Switch Fabric B are selected in the order [ABAB . . .] for transactions 302-305.

[031] Therefore, each destination (or receiving) IOP module expects that the traffic sequence from each source (or transmitting) IOP module will follow a sequence [ABABAB . . .] for a two switch fabric configuration. Similarly, in a three switch fabric router, each receiving IOP module expects that the traffic sequence from each transmitting IOP module will follow a sequence [ABCABCABC . . .]. Likewise, in a four switch fabric router, each receiving IOP module expects that the traffic sequence from each transmitting IOP module will follow a sequence [ABCDABCD . . .]. It should be noted that the sequence may begin anywhere, so that a [BABABA . . .] sequence is considered identical to [ABABAB . .]. The destination IOP module forwards data packets out the network interface ports alternately from each switch fabric.

[032] If all packets were identical in length, as in the case of a cell-based system, and input port contentions were minimized or eliminated, router 100 could use this behavior to mitigate the

mis-sequencing of data packets. However, variable packet lengths (e.g., between 64 and 1524 bytes) and the probability of input contention on any given IOP switch fabric interface result in significant packet variability. In FIGURE 3, exemplary packet sizes are shown for each data packet sent to IOP 5 from IOP 3. In this case, a large packet (1500 bytes) is sent among several small packets. The second packet through switch A (transaction 304) should arrive at IOP 5 prior to the first packet through switch B (transaction 303). In this case, the second packet from Switch A must be buffered while awaiting the first packet from switch B, so that packet order is maintained on the packets output to the network ports.

[033] Using a 1 Gigabit Ethernet switch fabric as an example, the packet delay would range between .512 microseconds and 12.192 microseconds. Assuming a worst case condition in which a maximum length packet is sent between two minimum length packets in the face of asymmetric input port congestion at the destination IOP module, a delay of $3\sigma(P_{1max} - P_{1min})$ can be assumed, where P_{1max} and P_{1min} are the maximum and minimum packet sizes and the three sigma (3σ) point of the traffic distribution is assumed. Thus, a delay of around 35 microseconds, or about 70 minimum length packets, could reasonably be assumed.

[034] In addition to providing a mechanism for sequencing packets across a load-shared switch fabric, the present invention can provide failure detection in the switch fabric path between the source IOP module and the destination IOP module. If the receiving IOP module detects a sequence failure, the assumption is that it is the result of a switch fabric failure or a failure of the associated optics. When a receiving IOP module detects the loss of traffic on a particular channel, it stops sending to the source IOP module on the faulty switch fabric and forwards all traffic to that source IOP module via the remaining switch fabrics. For example, the loss of traffic from Switch Fabric A for a specified interval causes the receiving IOP module to stop forwarding traffic on Switch Fabric A and to forward all of the traffic to that IOP module via Switch Fabric B. The source IOP module will then independently decide that Switch Fabric A is faulty.

[035] According to the principles of the present invention, the source and destination IOP module pairs use redundant switch fabrics in a round robin fashion to provide a mechanism for maintaining packet sequencing through the distributed architecture of router 100. Advantageously, this scheme requires packet buffering of a reasonably small size in the IOP modules. Thus, router 100 permits multiple switch fabrics to be used in a load-

sharing manner. Router 100 detects any departure from the normal round-robin sequencing, thereby permitting the failed switch fabric routes to be dropped.

[036] Although the present invention has been described with an exemplary embodiment, various changes and modifications may be suggested to one skilled in the art. The present invention is intended to encompass such changes and modifications as fall within the scope of the appended claims.